Artificial Intelligence Software Competitions: Boon or Bane ?

Moderator: Steve Chien, Jet Propulsion Laboratory

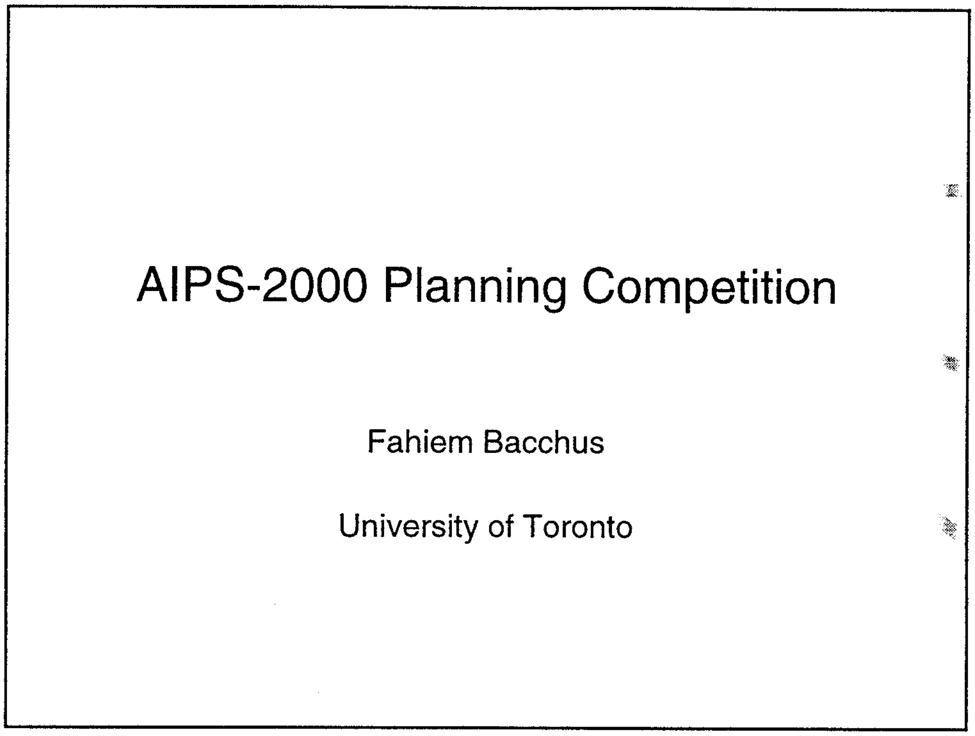
Planning: Fahiem Bacchus, University of Toronto

Automated Deduction: Geoff Sutcliffe, University of Miami

Trading Agents : Mike Wellman, University of Michigan

Ground Rules

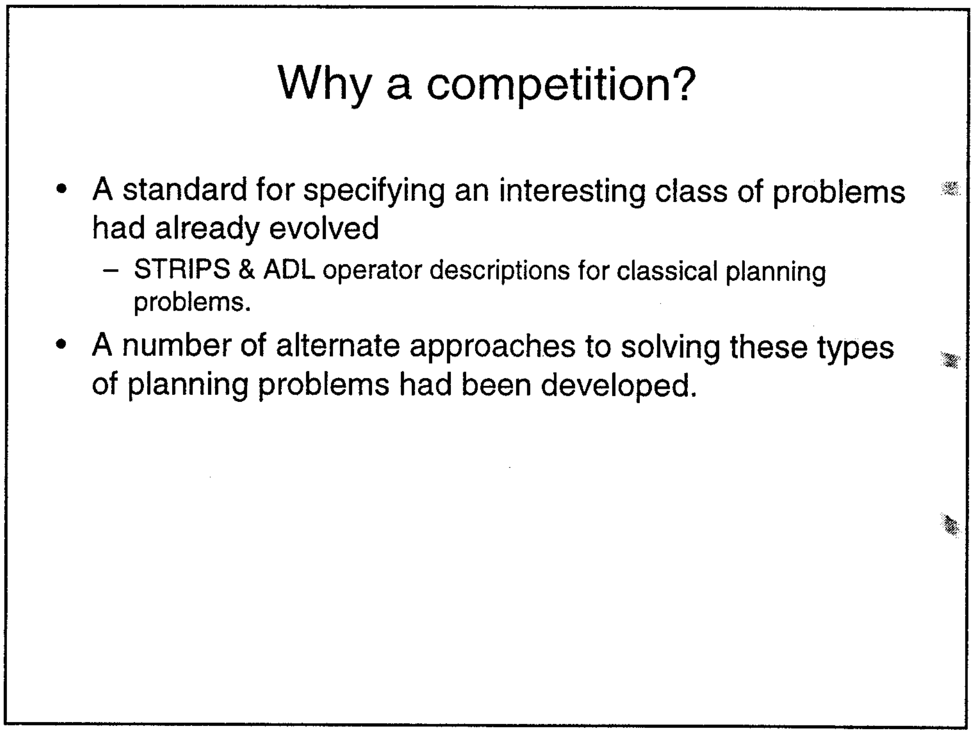
- Individual presentations on 3 competitions
- Brief discussions of individual competition
  - Background
  - Pros and Cons
- General discussion of Competitions
  - Audience participation!



AIPS-2000 Planning Competition

Fahiem Bacchus

University of Toronto



Why a competition?

- A standard for specifying an interesting class of problems had already evolved
  - STRIPS & ADL operator descriptions for classical planning problems.
- A number of alternate approaches to solving these types of planning problems had been developed.

Why a competition…

- Empirical evaluation had become the norm.
  - New approaches are generally implemented and tested on various problem suites.
- Major advances in planning performance, and computer hardware.
  - Inexpensive computers can be used to test these systems on complex problems.

The 2000 Competition

- The first competition held in 1998.
  - Two tracks based on the expressiveness of the operator: STRIPS & ADL.
  - 5 Competitors.
- The 2000 competition was the second.
  - Two tracks based on the amount of domain knowledge utilized: operator descriptions only or additional domain specific knowledge.
  - 15 Competitors.
  - 9 distinct approaches.
  - 8 Countries.

The 2000 Competition

- Track 1. The systems could take as input only the domain operators, the initial state, and goal of the particular problem instance.
  - The classical approach to solving the planning problem.
- Track 2. The systems could utilize additional domain specific information.
  - E.g., the HTN approach to planning.

Competitors did not have to use AI!

- In Track 2 no restrictions were placed on the kind of domain knowledge that could be utilized.
- A specific program could have been written for each domain!

Ultimately AI technology must show itself to be more cost effective than writing domain specific programs. So it seems legitimate to allow domain specific programs into the competition---as long as development costs are accounted for.

Benefits of the Competition

- Generates excitement, publicity, and interest in the field.
- Put proposed algorithms and approaches to a much more severe test---can these approaches make it to the next level?
- Can provide more informative and unbiased empirical evidence about the performance of different approaches.
- If done properly can help push research forward.

Disadvantages of the Competition

- Evaluation of research contribution might become too heavily dependent on performance in the competition.
  - The field must strive to maintain a broad and far reaching vision. The importance of different types of research must be recognized. We cannot be narrow minded about the merit of different types of research contribution...not everything needs to be immediately applicable or even remotely applicable to improving performance!

Disadvantages of the Competition

- Resources might be diverted from making fundamental advances to worrying about implementation details.
  - The competition must be organized in such a way that it forces people to focus on new problems at the frontier of research rather than on problems which require fine tuning existing techniques.

Summary

If used properly competitions can be very beneficial to a field.

I believe that with care they can be used properly.

Automated Deduction

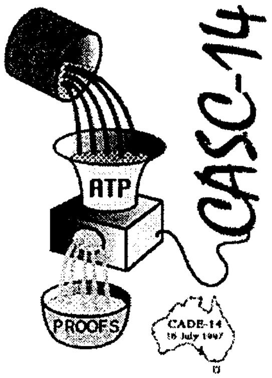
Geoff Sutcliffe

University of Miami

History

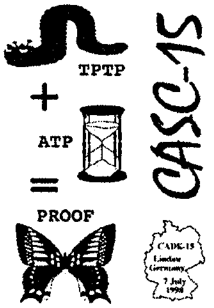
- Conceived in a park in Nancy, after CADE-12
- ATP community had "never [been] able to formulate an acceptable mechanism for comparing different systems" [Overbeek]
- A difficult and arguable venture
- Some inevitable constraints, some decisions
- First CASC at CADE-13
- Now established as an influential event in ATP

Organization

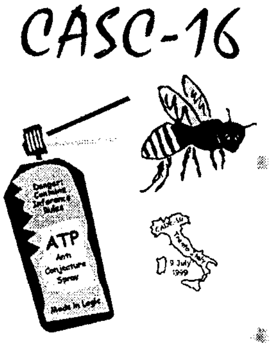


- Held annually at CADE
  - CASC-JC - Siena, Italy
  - CASC-17 - CMU, USA
  - CASC-16 - Trento, Italy
  - CASC-15 - Lindau, Germany
  - CASC-14 - Townsville, Australia
  - CASC-13 - Rutgers, USA
- Overseen by a panel of three knowledgeable researchers. Panel members have been ...
- Online design and rules, in advance
- Great T-shirts

Systems, Problems, and Ranking



- Systems
  - Classical 1st order logic
  - Sound, may be Incomplete, Automatic
  - Assurances and Solutions
  - Examples: Otter, E, Vampire, MACE, Waldmeister.
- Problems
  - Unbiased TPTP rated 0.21-0.99
  - Use unseen problems to negate tuning
- Ranking



Divisions

- MIX

```
input_clause(solve,conjecture,
    [--g(A,B),
     --g(B,a)]).
```

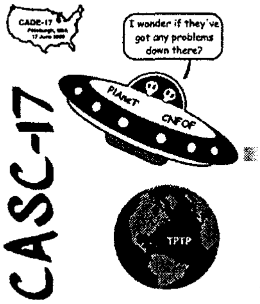
- UEQ

```
input_clause(left_inverse,axiom,
    [++equal(multiply(inverse(Y),Y,X),X)])
```

- FOF

```
input_formula(pel59_1,axiom,(
  ! [X] :
    ( big_f(X)
  <=> ~ big_f(f(X)) )    )).
```

- SAT
    - EPR



Rules for Entry

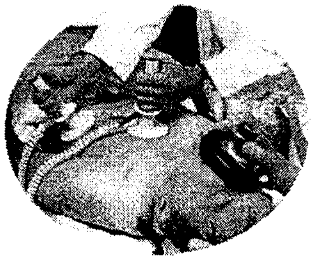
- ATP systems can be entered at only the division level
- ATP systems can be entered into more than one division
- A system that is not entered into a particular division is assumed to perform worse than the entered systems
- Previous winner automatically entered in each division
- Robust installation
- Automatic and clean operation
- No storing information for individual TPTP problems
- Repeatable operation
- Soundness testing before and after competition
- Solutions checked before and after competition
- Sources made publically available on WWW after competition

- Catch-all rule: No cheating is allowed

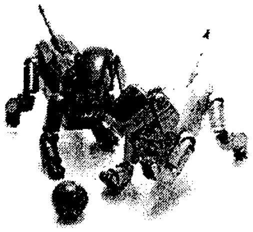
Results

	MIX	UEQ	SAT	FOF	EPR
CASC-JC	E-SETHEO VampireJC	Waldm'r	-	E-SETHEO	E-SETHEO
CASC-17	E	Waldm'r	Gandalf	Vampire	
CASC-16	Vampire	Waldm'r	MACE	SPASS	
CASC-15	Gandalf	Waldm'r	SPASS	SPASS	
CASC-14	Gandalf	Waldm'r	SPASS	SPASS	
CASC-13	E-SETHEO	Otter			

Pros



Stimulates
ATP research
in general

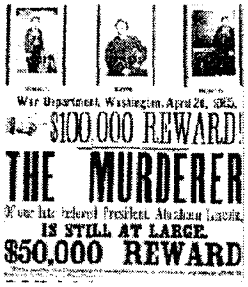


Stimulates research towards
autonomous systems

Provides motivation for
implementing and fixing
systems

Pros

Evaluates ATP systems



War Department, Washington, April 20, 1865,

$100,000 REWARD!

THE MURDERER

Of our late beloved President, Abraham Lincoln,

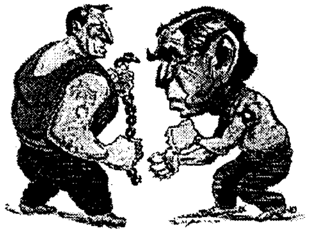
IS STILL AT LARGE.

$50,000 REWARD

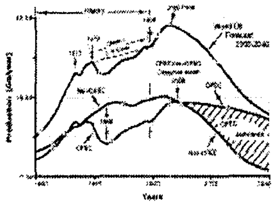
Rewards implementation efforts

Exposes ATP systems within and outside the ATP community

Pros



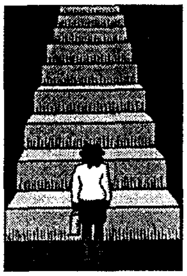
Provides an inspiring environment for personal interaction between ATP researchers



Tests the theory

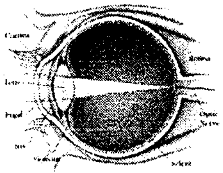
Inspires other (deduction) competitions

Cons

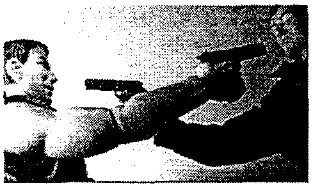


Incremental development

Excessive tuning



Focus on implementation at cost of theory



Tension

Trading Agent Competition

Mike Wellman

University of Michigan

Trading Agent Competition

- "Open invitation" market game
- TAC 2000
  - Attracted 20 entries from 6 countries
  - Held July 2000, Boston, at ICMAS-00
- TAC 2001 to be held in October, Tampa
- …focus on trading strategies, implications for automated commerce

http://tac.eecs.umich.edu/

TAC Travel Shopping Game

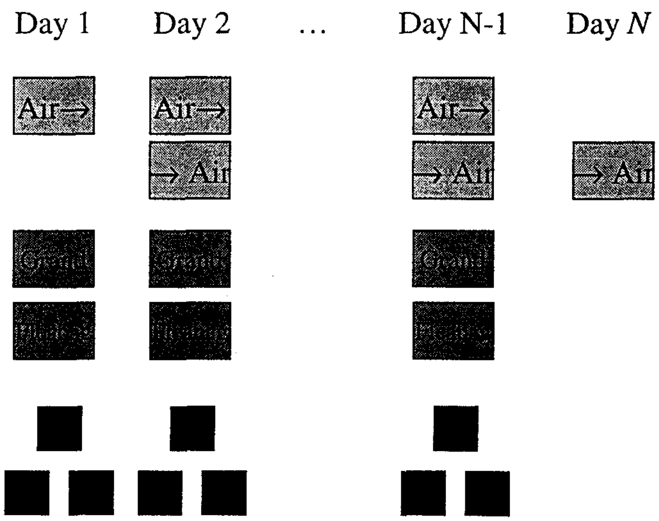
- Each travel agent given clients' requests, defining objective function. Net value is objective minus expenditure.
- Assemble trip for each client, comprising flight, hotel, and entertainment.
- Goods are interdependent, each presents interesting issues.

Agent Objectives

- Maximize total “profit”:

  [sum over clients: trip utility] minus expenditures
- Client preferences: arrive/depart days, hotel premium, entertainment prefs
- Feasible trip: round trip airline, hotel room for interval
- Trip utility: zero if infeasible
  - if feasible... 1000 – date deviation penalty + premium hotel bonus + entertainment bonus

Auction Configuration



Key Game Insights

- Be robust to server/network conditions.
- Value least commitment.
- Hotels are bottleneck resource.
- Ignore sunk costs.
- Model aggregate, but not individual, agent behavior.
- Global optimization (vs. client-by-client).

TAC Future

- Analysis and experimentation continues.
- Maintain and package for use in courses.
- Rule revisions:
  - make flights more interesting
  - ameliorate hotel "witching hour"
  - distribute goods to promote more trade
- Planning for a bigger and better TAC-01.
  - At EC-01, 14 Oct 01 in Tampa
  - Still time to sign up! (or sponsor...)

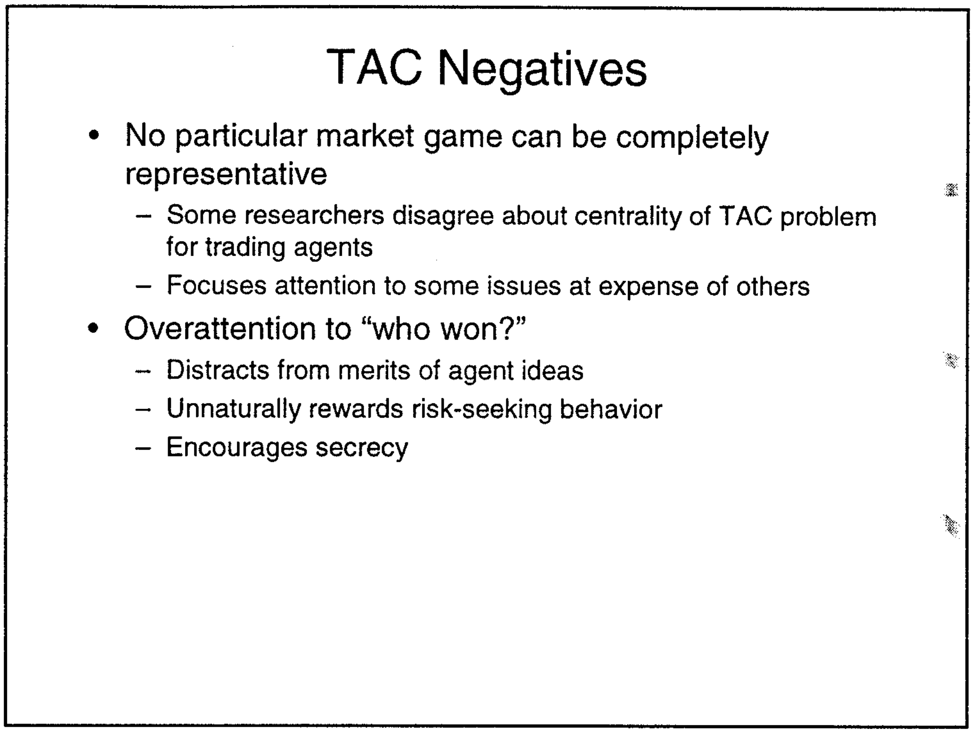
http://tac.eecs.umich.edu/

TAC Positives

- Focus effort on common problem
  - Previous trading-agent work hard to compare
  - Lack of discussion about what is the relevant problem, performance metrics
- Source of agent trading strategies
  - Cannot judge a strategy in isolation, so research in the area requires model of *other* agents
  - Where do such models come from?
- Encourage openness
  - Commercial "programmed traders" very secretive

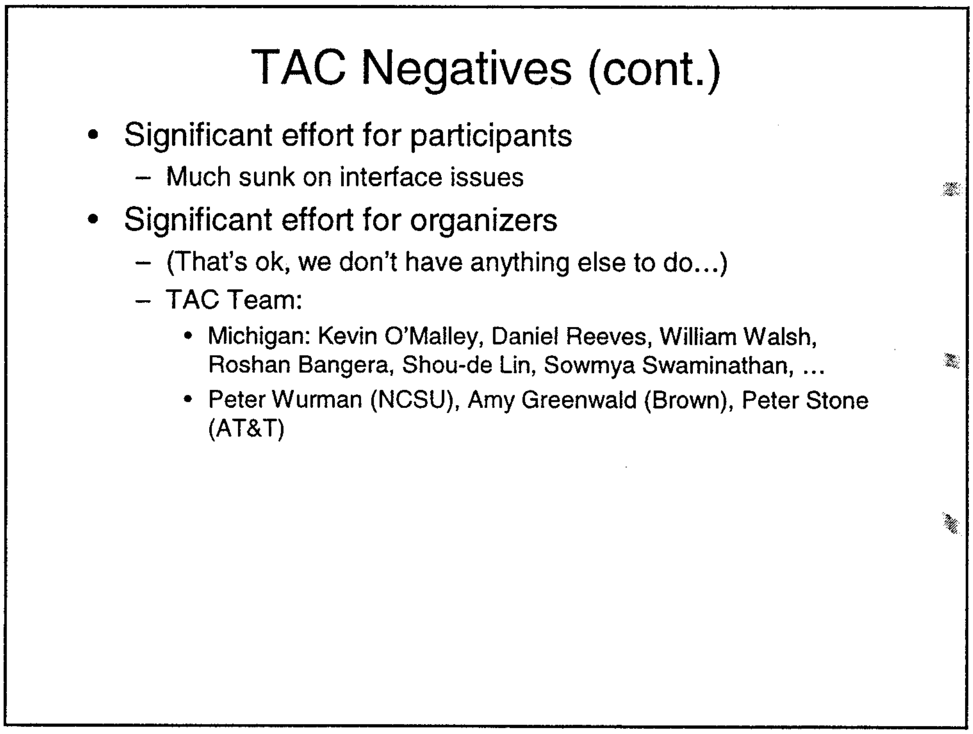
TAC Positives (cont.)

- Generate excitement
  - Within research community
  - Related communities, general public
  - Entry point for nontraditional researchers
- Publications (by entrants and organizers)
  - IEEE Internet Computing (2), JAIR
  - Agents-01, EC-01
  - Dr. Dobb's
  - Popular press: NYT, others



TAC Negatives

- No particular market game can be completely representative
  - Some researchers disagree about centrality of TAC problem for trading agents
  - Focuses attention to some issues at expense of others
- Overattention to “who won?”
  - Distracts from merits of agent ideas
  - Unnaturally rewards risk-seeking behavior
  - Encourages secrecy



TAC Negatives (cont.)

- Significant effort for participants
  - Much sunk on interface issues
- Significant effort for organizers
  - (That’s ok, we don’t have anything else to do…)
  - TAC Team:
    - Michigan: Kevin O’Malley, Daniel Reeves, William Walsh, Roshan Bangera, Shou-de Lin, Sowmya Swaminathan, …
    - Peter Wurman (NCSU), Amy Greenwald (Brown), Peter Stone (AT&T)

Questions?

- Should we have more competitions?
- Should we have papers selected by competition?
- Should we have funding sources selected by competition?
- Are the competition metrics the important ones?
  - Knowledge acquisition / implementation v. runtime?
  - Analogy to Knowledge Acquisition competitions?
- What are the relative merits of benchmark problems versus actual competitions (benchmarks = slow competitions)
- How can we encourage development of realistic benchmarks and domains?
- What are examples of innovations (algorithms, approaches?) that can be traced to competitions?
  - If none, why have them?
  - Competitions rules → all systems from each year must be released to public by next year?